Compatibility, Cliques and Clonal Frames

Barbara Holland
University of Tasmania
Unravelling the processes of bacterial evolution

• Processes
  – Mutation
  – Homologous recombination
  – HGT

• Data is available at multiple levels of resolution
  – Gene presence / absence
  – Allele profile
  – Sequence data
Compatibility

Given a character C and a tree T we can ask if the character is compatible with the tree.
Compatibility

Given a character C and a tree T we can ask if the character is compatible with the tree.

A compatible character
Incompatibility
Incompatibility

An incompatible character
Compatible cliques of characters

• Characters are said to be compatible with each other if there exists a tree which they are all compatible with.
Allele profile data

- Multi-level data
  - Strain type
  - Allele profile
  - Sequence

<table>
<thead>
<tr>
<th>Locus</th>
<th>L1</th>
<th>L2</th>
<th>L3</th>
<th>L4</th>
<th>L5</th>
<th>L6</th>
<th>L7</th>
</tr>
</thead>
<tbody>
<tr>
<td>ST1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ST2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>....</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

e.g. MLST data

L3

<table>
<thead>
<tr>
<th></th>
<th>CCCTTGTTCAGTCCAAATTCACACCAATTTCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Allele profile data

- Multi-level data
  - Strain type
  - Allele profile
  - Sequence

e.g. MLST data

<table>
<thead>
<tr>
<th>locus</th>
<th>L1</th>
<th>L2</th>
<th>L3</th>
<th>L4</th>
<th>L5</th>
<th>L6</th>
<th>L7</th>
</tr>
</thead>
<tbody>
<tr>
<td>ST1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ST2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>....</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

L3

1

CCCTTGTTTAGTCCAAATTCACACCAATTTCA

2

CCCTTATCTGGCTCAAATTTCACACCAATTTCA

...

...

Evolution of a single **locus** along a clonal frame by mutation (M) and recombination (R) events. A locus is a contiguous stretch of DNA – it will be represented by one column in an allele profile.

Allele types

1. `ACCGATATAGGATCGTTCGTCA`
2. `ACCGTTGCAGGACTGCTAGCCA`
3. `ACCGTTGCAGGTCTGCTAGCCA`

Allele type 2 and 3 differ from each other in a single position due to a mutation event. Allele type 1 and 2 differ from each other in many positions due to a recombination event. This locus makes up a single column (bold) of the allele profile below.

<table>
<thead>
<tr>
<th>Allele Profile</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
</tr>
<tr>
<td>B</td>
</tr>
<tr>
<td>C</td>
</tr>
<tr>
<td>D</td>
</tr>
<tr>
<td>E</td>
</tr>
</tbody>
</table>
A range of recombination models

(A) ClonalFrame model: Recombination always introduces novel genetic material.

(B) Intermediate model

(C) ClonalOrigin model: Recombination always occurs within a closed population.

Open system  Closed system
A particular locus can undergo two types of events: mutation and recombination.

Parallel mutation should be infrequent as it requires
1) that the next mutation in the sequence for that locus occurs at the same site, i.e. without any other mutations occurring in the meantime $p \propto \frac{1}{L}$
2) And it further requires that the mutation is back to the initial state

Parallel recombination might be more likely, especially in a closed system. In an open system – as per the ClonalFrame model – parallel recombination should be even less likely than parallel mutation.
Loci that haven’t undergone parallel recombination will produce a character (i.e. a column in the allele profile) that is compatible with the clonal frame.

Blocks that have undergone parallel recombination (or parallel mutation) may produce characters that are not compatible with the clonal frame.
The *Campylobacter jejuni* data

- 46 *C. jejuni* genomes
- 686 genes in common across all 46 genomes
Initial analysis

• 686 characters
• 9 constant, 2 parsimony uninformative
• Theoretical best parsimony score 7083

\[ \sum_{l=1}^{686} (n_l - 1) \]

Where \( r_l \) is the number of alleles at locus \( l \)

• Parsimony finds 3 equally parsimonious trees with score 8274
• Consistency index 0.856
Tree is well supported

Consensus network of 100 parsimony btsp trees showing splits with > 20% support

Edge length proportional to support
Are some genes more prone to parallel events?

Under the infinite alleles model all characters should have excess 0.

Here there were 213 compatible (excess 0) characters

And 473 characters that required at least 1 extra mutation
Ancestral state reconstruction

• Find the clonal frame using maximum parsimony

• Use parsimony version of ASR work out all the transitions from one allele to another – look at the distribution of differences between pairs of alleles.

• Compare the distribution of allele differences of compatible characters to that of incompatible characters
Clear cases of parallel recombination

 Allele 0 and 1 differ at 20 sites

100185noOut.fa
18 alleles
Excess of 3
Are parallel events more often mutation or recombination?

Relative frequencies of allele differences

Number of differences between alleles

- Parallel Changes
- All Changes
Are some edges more prone to recombination events?

• See scribbles
Tree rooted fairly arbitrarily (multipoint attempt by eye)
Edges lengths not to scale in this picture

Edges annotated with
allele-distance / gene-name

Top 15 shown on tree
largest allele distances

on edges means > 30%
events have
> 10 allele distance

on edges means > 5%
events have > 20
-> 501 parallel events

Events shown at top 15
Are some edges or clades more prone to parallel events?

Bold edges / labels indicate that more than 50% of events allocated to that edge are parallel events.

Retention index by clade (#parsimony inf. characters)
Conclusions

• Overall AP data is very consistent, i.e. highly compatible, consistency index > 0.85

• Clonal Frame wastes a lot of computational effort on finding the clonal frame but its model predicts (close to) perfect phylogenies.

• Hard to tell if parallel mutation is more common than parallel recombination as recombination might occur frequently between alleles that aren’t very different.

• Seems like different processes predominate in different parts of the tree. Sampling artefact? Testable?
Acknowledgements

• Nigel French, Patrick Biggs, Shoukai Yu

• Marsden Fund grant to NF